# CHAPTER 15
# RELIABILITY

Although an individual question's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way that questions function together and complement one another. Any measurement includes some amount of measurement error; that is, no measurement can be perfectly accurate. This is true of academic assessments—no assessment can measure students perfectly accurately; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. Questions that function well together produce assessments that have less measurement error; that is, the errors made should be small on average. Such assessments are described as reliable.

There are a number of ways to estimate an assessment's reliability. One approach is to split all test questions into two groups and then correlate students' scores on the two half tests. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, questions on the two half tests must be measuring very similar knowledge or skills. This is evidence that the questions complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires the psychometrician to select which questions contribute to each half-test score. This decision may have an impact on the resulting correlation. Cronbach (1951) provided a statistic that avoids this concern about the split-half method. Cronbach's $\alpha$ coefficient is an estimate of the average of all possible split-half reliability coefficients.

## RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

Table 15-1 presents descriptive statistics, Cronbach's $\alpha$ coefficient, and raw and scaled score standard errors of measurement for each subject separately for each grade level. The reported reliability for Writing, Health, and Visual and Performing Arts are the averages of the computed Cronbach's $\alpha$ across forms. The low reliability values can be attributed to the lower number of items in each form in those tests.

Note, two scaled-score standard errors of measurement are presented: one for scaled scores below 542 and one for scaled scores of 542 and above. This is because different slopes are used in the linear transformation to scaled scores at these two different parts of the scaled score range.

Table 15-1
Reliabilities, Standard Errors of Measurement and Descriptive Statistics

| Grade | Subject | n | Raw Score | | | | | | Scaled Score | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | <542 | >=542 |
| | | | Min. | Max. | Mean | S.D. | Rel. | S.E.M. | S.E.M. | S.E.M. |
| 4 | Reading | 15,546 | 0 | 52 | 32.39 | 7.47 | .84 | 3.01 | 3.60 | 2.20 |
| | Writing | 15,704 | 6 | 28 | 14.66 | 4.64 | .69 | 2.60 | 4.09 | 2.73 |
| | Mathematics | 15,838 | 0 | 40 | 20.50 | 7.43 | .80 | 3.36 | 4.79 | 2.62 |
| | Science | 15,898 | 0 | 36 | 16.59 | 5.57 | .74 | 2.86 | 3.90 | 2.75 |
| | Social Studies | 15,885 | 0 | 37 | 22.79 | 4.63 | .70 | 2.54 | 5.07 | 2.94 |
| | Health | 15,896 | 0 | 18 | 11.47 | 2.31 | .49 | 1.65 | 3.95 | 4.66 |
| | Visual and Performing Arts | 15,935 | 0 | 18 | 11.47 | 2.43 | .45 | 1.80 | 7.24 | 5.58 |
| 8 | Reading | 16,347 | 0 | 51 | 31.42 | 7.37 | .83 | 3.08 | 3.90 | 2.73 |
| | Writing | 16,385 | 6 | 29 | 16.01 | 4.47 | .68 | 2.52 | 4.88 | 3.21 |
| | Mathematics | 16,337 | 0 | 41 | 16.78 | 9.03 | .85 | 3.52 | 3.83 | 2.44 |
| | Science | 16,424 | 0 | 38 | 19.33 | 5.48 | .74 | 2.82 | 4.71 | 2.86 |
| | Social Studies | 16,353 | 1 | 38 | 19.14 | 5.15 | .73 | 2.66 | 5.13 | 3.20 |
| | Health | 16,377 | 0 | 22 | 12.55 | 3.27 | .50 | 2.31 | 4.04 | 3.56 |
| | Visual and Performing Arts | 16,467 | 0 | 22 | 12.81 | 3.55 | .54 | 2.41 | 6.48 | 3.82 |
| 11 | Reading | 13,501 | 0 | 52 | 35.90 | 7.30 | .84 | 2.93 | 4.10 | 2.40 |
| | Writing | 13,638 | 7 | 28 | 17.73 | 4.14 | .68 | 2.32 | 4.63 | 3.19 |
| | Mathematics | 13,398 | 0 | 41 | 16.67 | 5.57 | .70 | 3.04 | 5.53 | 3.63 |
| | Science | 13,496 | 0 | 38 | 16.93 | 4.75 | .59 | 3.02 | 5.81 | 3.35 |
| | Social Studies | 13,392 | 0 | 37 | 16.43 | 3.77 | .55 | 2.52 | 7.16 | 3.92 |
| | Health | 13,416 | 0 | 22 | 13.19 | 3.40 | .47 | 2.48 | 4.48 | 3.84 |
| | Visual and Performing Arts | 13,484 | 0 | 22 | 13.43 | 3.59 | .50 | 2.53 | 7.81 | 4.15 |

*The reported reliability is the average reliability across forms.

## STRATIFIED COEFFICIENT ALPHA

According to Feldt and Brennan (1989) a prescribed distribution of items over categories (such as different item types) indicates the presumption that at least a small, but important, degree of unique variance is associated with the categories. In contrast, Cronbach's coefficient $\alpha$ is built upon the assumption that there are no such local or clustered dependencies. A stratified version of coefficient $\alpha$ corrects for this problem.

Stratified coefficient α was calculated separately for each common item test and grade level. The stratification was based on item types (multiple-choice v. open response). These results are provided in Table 15-2.

Table 15-2
Coefficients α and Stratified α

| Grade | Content | α | $\alpha_{mc}$ | $N_{mc}$ | $\alpha_{or}$ | $N_{or}$ (Pts.) | Stratified α |
|---|---|---|---|---|---|---|---|
| 4 | Reading | 0.87 | 0.79 | 23 | 0.77 | 11 (30) | 0.87 |
| | Mathematics | 0.80 | 0.72 | 15 | 0.70 | 8 (26) | 0.82 |
| | Social Studies | 0.74 | 0.60 | 15 | 0.66 | 7 (24) | 0.75 |
| | Science | 0.74 | 0.57 | 15 | 0.63 | 8 (26) | 0.75 |
| 8 | Reading | 0.86 | 0.76 | 23 | 0.78 | 11 (30) | 0.86 |
| | Mathematics | 0.85 | 0.77 | 15 | 0.80 | 7 (26) | 0.88 |
| | Social Studies | 0.79 | 0.65 | 15 | 0.74 | 8 (26) | 0.81 |
| | Science | 0.77 | 0.58 | 15 | 0.73 | 8 (26) | 0.78 |
| 11 | Reading | 0.89 | 0.82 | 23 | 0.84 | 11 (30) | 0.90 |
| | Mathematics | 0.84 | 0.72 | 15 | 0.81 | 7 (26) | 0.87 |
| | Social Studies | 0.80 | 0.58 | 22 | 0.80 | 6 (24) | 0.83 |
| | Science | 0.77 | 0.60 | 15 | 0.71 | 7 (26) | 0.79 |

## RELIABILITY OF PERFORMANCE LEVEL CATEGORIZATION

All test scores contain measurement error; thus classifications based on test scores are also subject to measurement error. After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications.

### ACCURACY

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

### CONSISTENCY

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel, form of the same test. Consistency can be evaluated directly from actual responses to test questions if two complete, parallel, forms of the test are given to the same group of students. This is usually impractical, especially on lengthy tests such as the MEA. To overcome this issue, techniques have been developed to estimate both accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the MEA because their technique can be used with both constructed-response and multiple-choice questions.

## CALCULATING ACCURACY

All of the accuracy and consistency estimation techniques described below make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. Following Livingston and Lewis (1995), the true-score distribution for the MEA was estimated using a four-parameter beta distribution, which is a flexible model that allows for extreme degrees of skewness in test scores.

In the Livingston and Lewis method, the estimated "true scores" are used to classify students into their "true" performance category, which is labeled "true status." After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4 × 4 contingency table is created for each test and grade level. The cells in the table are the proportion of students who were classified into each performance category by the actual (or observed) scores on the MEA (i.e., observed status) and by the "true scores" (i.e., "true status"). As an example, Table 15-3 shows the accuracy contingency table for fourth-grade Social Studies. The accuracy contingency tables for all grades and subjects are provided in Appendix D (under step 5). Additional steps in the analysis are also shown in Appendix D.

Table 15-3
Accuracy Contingency Table for Grade 4 Social Studies

| True Status | Observed Status | | | |
|---|---|---|---|---|
| | Does Not Meet the Standards | Partially Meets the Standards | Meets the Standards | Exceeds the Standards |
| Does Not Meet the Standards | **.06** | .03 | .00 | .00 |
| Partially Meets the Standards | .06 | **.49** | .09 | .00 |
| Meets the Standards | .00 | .05 | **.19** | .00 |
| Exceeds the Standards | .00 | .00 | .01 | **.01** |

Proportions on the diagonal (in bold) indicate exact agreement between the observed status and "true status." If the test were perfectly accurate, all of the off-diagonal cells would be zero. Accuracy is the sum of the diagonal (i.e., the proportion of exact agreement across the four performance levels). In Table 15-3, the diagonal sums to .75, indicating that 75 percent of the students were classified into exactly the same performance categories by their observed scores and their "true scores."

### KAPPA

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classification that would be expected by chance. Cohen's κ can be used to estimate the classification consistency of a test from two parallel forms of the test. The second form in this case was the one estimated using the Livingston and Lewis (1995) method. Cohen's κ is shown in Table 15-4. Because κ is corrected for chance, the values of κ are lower than the other consistency estimates in Table 15-4.

### CALCULATING CONSISTENCY

To estimate consistency, the "true scores" are used to estimate the distribution of classifications on an independent, parallel test form. After statistical adjustments (see Livingston and Lewis, 1995), a new $4 \times 4$ contingency table is created for each test and grade level that shows the proportions of students who were classified into each performance category by the actual test and by another (hypothetical) parallel test form. Consistency, which is the proportion of students classified into exactly the same categories by the two forms of the test, is the sum of the diagonal for the new contingency table. The consistency contingency tables are shown under step 7 in Appendix D.

### RESULTS OF ACCURACY, CONSISTENCY, AND KAPPA ANALYSES

The accuracy, consistency, and kappa indices for all grades and subjects are summarized in Table 15-4.

| Table 15-4<br>Estimates of Accuracy and Consistency of Performance Level Classification | | | | |
|---|---|---|---|---|
| Grade | Subject | Accuracy | Consistency | Kappa (κ) |
| 4 | Reading | 0.8 | 0.74 | .55 |
| | Writing | 0.86 | 0.79 | .63 |
| | Mathematics | 0.76 | 0.67 | .46 |
| | Science | 0.81 | 0.72 | .44 |
| | Social Studies | 0.75 | 0.65 | .39 |
| | Health | 0.74 | 0.63 | .26 |
| | Visual and Performing Arts | 0.64 | 0.52 | .22 |
| 8 | Reading | 0.8 | 0.72 | .54 |
| | Writing | 0.85 | 0.79 | .63 |
| | Mathematics | 0.79 | 0.71 | .55 |
| | Science | 0.77 | 0.67 | .44 |
| | Social Studies | 0.76 | 0.66 | .45 |
| | Health | 0.73 | 0.64 | .27 |
| | Visual and Performing Arts | 0.69 | 0.57 | .30 |
| 11 | Reading | 0.8 | 0.74 | .57 |
| | Writing | 0.83 | 0.77 | .61 |
| | Mathematics | 0.78 | 0.69 | .52 |
| | Science | 0.8 | 0.72 | .47 |
| | Social Studies | 0.75 | 0.65 | .44 |
| | Health | 0.72 | 0.61 | .25 |
| | Visual and Performing Arts | 0.68 | 0.55 | .31 |

For certain decisions, concern may be highest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of four or five, but not one, two, or three, one might be interested in the accuracy of the dichotomous decision, below four versus four or above. Table 15-5 reports accuracy and consistency for various dichotomous categorizations on the

MEA. MEA P/M cut accuracy ranges from .77 to .97, and M/E accuracy ranges from .97 to .999. These are relatively high values compared to the 1999 Advanced Placement (AP) accuracy of decisions based on the 2-3 cut and 3-4 cut which ranges from .84 to .95.

Table 15-5

Accuracy and Consistency of Dichotomous Categorizations

| Grade | Subject | Accuracy | | | Consistency | | |
|---|---|---|---|---|---|---|---|
| | | D/P | P/M | M/E | D/P | P/M | M/E |
| 4 | Reading | .93 | .88 | .98 | .92 | .84 | .98 |
| | Writing | .92 | .94 | .99+ | .88 | .91 | .99+ |
| | Mathematics | .88 | .89 | .99 | .83 | .84 | .98 |
| | Science | .84 | .97 | .99+ | .76 | .95 | .99+ |
| | Social Studies | .90 | .86 | .99 | .86 | .80 | .98 |
| | Health | .97 | .78 | .98 | .95 | .69 | .98 |
| | Visual and Performing Arts | .81 | .85 | .98 | .74 | .78 | .97 |
| 8 | Reading | .93 | .89 | .98 | .91 | .84 | .97 |
| | Writing | .94 | .92 | .99+ | .91 | .86 | .99 |
| | Mathematics | .89 | .91 | .99 | .85 | .88 | .98 |
| | Science | .86 | .91 | .99+ | .79 | .88 | .99+ |
| | Social Studies | .90 | .88 | .99 | .85 | .83 | .99 |
| | Health | .97 | .77 | .99 | .96 | .68 | .99 |
| | Visual and Performing Arts | .85 | .84 | .99 | .79 | .79 | .98 |
| 10 | Reading | .93 | .89 | .98 | .92 | .85 | .97 |
| | Writing | .93 | .91 | .99 | .91 | .87 | .99 |
| | Mathematics | .88 | .91 | .99 | .83 | .87 | .98 |
| | Science | .86 | .94 | .99+ | .80 | .92 | .99+ |
| | Social Studies | .88 | .88 | .99 | .83 | .83 | .98 |
| | Health | .95 | .78 | .99 | .92 | .70 | .99 |
| | Visual and Performing Arts | .82 | .86 | .99 | .74 | .81 | .98 |